External Validation of Computational Barthel Index: Why Accuracy Drops?

Lemba Priscille Ngana, MPH, Janusz Wojtusiak, PhD
George Mason University, Fairfax, VA

Introduction

Computational Barthel Index (CBIT) has been developed to help clinicians, patients, and caregivers forecast upcoming changes to functional status of patients up to one year ahead.[1] It has been constructed using data from Community Living Centers (CLCs) in the Department of Veterans Affairs (VA). CBIT consists of 72 gradient boost models that predict functional dependency based on demographics, diagnoses and if available previous functional status. The models reported a high accuracy, with an average AUC of 0.94.[1] However, the authors of the original work did not report testing results on data from outside the VA, therefore potentially limiting the. As with many machine learning-based models, external validation is key to quantify the reproducibility and generalizability of models. Few studies conduct external validation, which often results in a change in model performance due to differences in the datasets (e.g., differences in distribution of top model predictors or outcome incidence).[2] This work aims at independently validating CBIT using data extracted from linked SEER-Medicare database for both cancer and non-cancer patients.

Methods

Data: SEER-Medicare 2006-2018 data were processed to exactly reproduce variables present in the original CBIT work[1]. The outcome variables have been created from Minimum Data Set (MDS) 2.0 and 3.0, by applying existing method[3] that creates indicators for activities of daily living: feeding, bathing, grooming, dressing, bowels, bladder, toilet use, transfers and mobility. The 50 input variables to CBIT included patient demographics and selected diagnosis codes (ICD-9 and ICD-10) mapped to CCS. For each diagnosis, the model encoded the number of days since the first known and the most recent occurrence of the code in patient records. Datasets have been created for testing and assessing current ADLs, as well as predicting them 90, 180 and 360 days ahead.

Data comparison: Mann-Whitney U test and Chi-squared test were used to determine significant differences between the two datasets ($p < 0.05$) on input and outcome variables.

Model evaluation: Model performance was tested using metrics used in the original CBIT work: area under receiver-operator curve (AUC), precision, recall and calibration. Data were stratified by gender, race, and age to test for differences in model performance in these sub-populations.

Results

The cohort included 843,836 patients and a total of 8,257,141 MDS evaluations completed between 2006 and 2018. Descriptive statistics indicated that patients in CMS dataset were predominantly female (60% in CMS vs. only 3% in VA), older in age (78 vs. 70 years old) and more disabled (83% vs. 63%) in comparison to the patients in the original VA dataset. Moreover, the average Barthel scores were 3.7 and 5.3 for patients in CMS and VA data, respectively. Of particular interest were the differences between the average CCS_max variables that indicate first occurrence of a condition (1101 days vs. 1312 days) and CCS_min variables that indicate the most recent occurrence (492 days vs. 571 days) values of the 2 datasets. Overall, a significant drop in CBIT model accuracy has been observed when applying CMS cohorts to CBIT gradient boost models, resulting in 69.3% AUC vs. 79.5% AUC obtained when using the VA dataset. The average performance of the models using CMS data versus VA data is as follows: AUC 59.5% vs. 76.1% at 3 months, 59.5% vs. 74.6% at 6 months and 59.2% vs. 72.5% at 1 year time point. Furthermore, greater model performances were observed in black race (68.7%), female gender (67.4%) and ages 60-70 years old (67.9%).

Conclusions

We hypothesized that the external validation of computational Barthel index using the SEER-Medicare dataset would lead to a decrease in the CBIT model performance. This hypothesis aligns with the results obtained and can be explained by significant differences and large variances between the two datasets ($p > 0.05$). Further work is needed to understand if the differences can be mitigated using transfer learning methods.

References

1. Wojtusiak J, Asadzadehzanjani N, Levy C, Alemi F, Williams AE. Computational Barthel Index: an automated tool for assessing and predicting activities of daily living among nursing home patients. BMC Med Inform Decis Mak. 2021;21(1):17.
2. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where?. Clin Kidney J. 2020;14(1):49-58.
3. Wojtusiak, J., Levy, C., Williams, A. and Alemi, F., "Predicting Functional Decline and Recovery following Hospitalization of Residents in Veterans Affairs Nursing Homes," The Gerontologist, 56(1), 2016.